

Life Sciences Identifier (LSID): A Foundation for Wide Area, Scientific Collaboration and Informatics Interoperability

August 27, 2004

Brian Gilman
Cold Spring Harbor Labs/Panther Informatics





Contents

- **Background**
 - Acknowledgements
 - The Big Ideas
 - Scientific Complexity and Informatics Interoperability
 - LSID Use Cases
 - LSID Requirements
- **How LSID Works**
 - Identifier Format and Semantics
 - LSID and other Identity Standards
- **Conclusions**
- **LSID Standard Submission Status**
- **Q & A**



Acknowledgements

- LSID Fellow Authors:
 - Ted Liefeld, Millennium Pharmaceuticals
 - Brian Gilman, The Whitehead Institute/Panther Informatics
 - Stephanos Bacon, Avaki Corporation
 - Josh Apgar, Avaki Corporation
 - Sean Marin, IBM Corporation

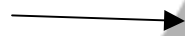
Life Sciences Identifier – The Big Ideas

- **LSID is an naming standard for distributed data, specifically:**
 - Biologically significant data items,
 - Located in distributed data stores,
 - Including files, database records, and data objects managed by N-tier applications,
 - That are accessible over public and/or private networks,
 - And owned, managed, and/or curated by different academic, research, government or commercial organizations.
- **LSID names are semantically void/opaque with respect to the objects they identify.**
- **LSID replaces physical addresses with opaque, location independent identifiers expressed as URNs.**
- **LSID complements Web services standards such as SOAP/XML, WSDL, UDDI, SAML, WS-Security, MS Passport, Liberty Alliance, GSS-API and OGSA.**

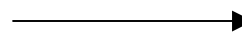
Complexity Requires Interoperability

Scientific Exploration

Genomics



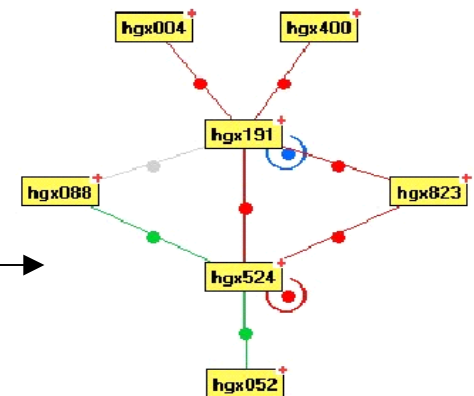
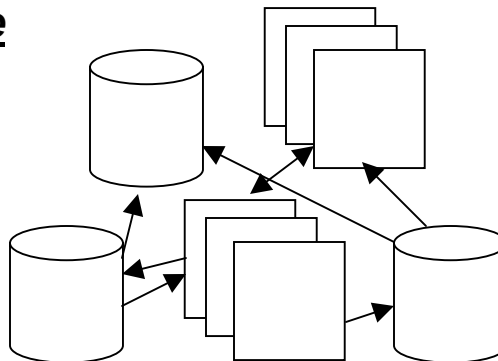
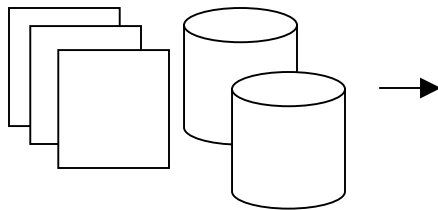
Proteomics



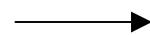
Pathways:
Regulation,
Signaling,
Transport

Multi-Discipline, Multi-Organizational Complexity

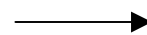
Informatics Infrastructure



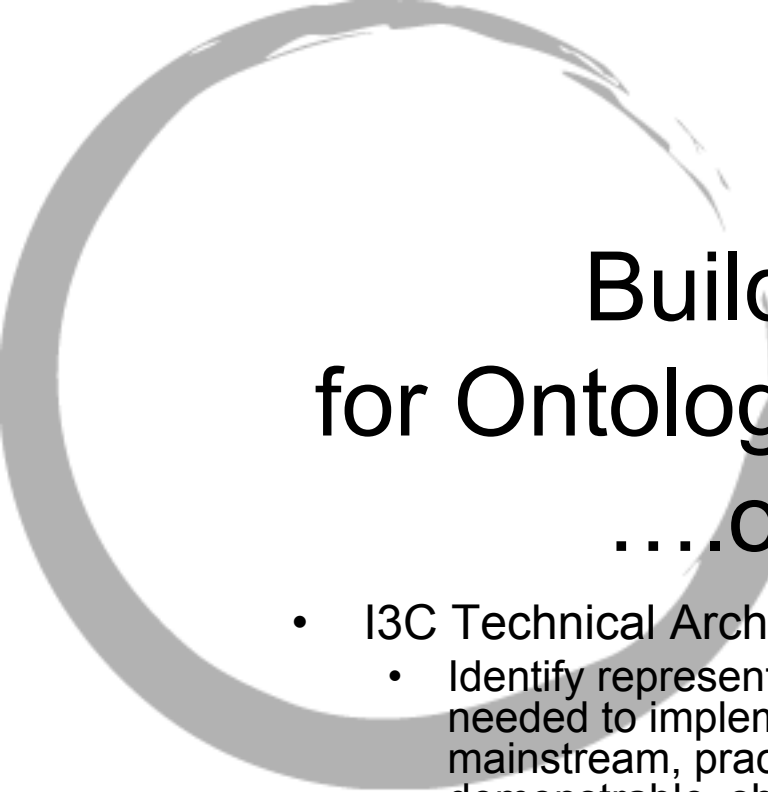
Data silos



Collaboration



Ontologies and Data Networks



Building a Foundation for Ontologies and Data Networks,one step at a time

- I3C Technical Architecture Group Manifesto
 - Identify representative use cases that anticipate the expressiveness needed to implement ontologies and data networks, while solving mainstream, practical informatics interoperability problems with demonstrable, short term, return on investment.
 - Continuously validate new tools and standards proposals with I3C discovery groups, such as I3C Pathways.
 - Test proposed standards by implementing prototypes that demonstrate real world use cases and rapidly iterate proposals.
 - Don't reinvent the wheel. Harness existing or emerging industry standards.
 - Drive mainstream adoption of enabling technology that is narrowly defined and easy to adopt throughout the life sciences community.
- 

LSID Use Cases

- Intranet Portal & Uniform Data Naming

Problem: Researchers need a common way to access internal databases, file systems, or n-tier applications using a standard web-based interface that let's them bookmark and refer back to many different types of data items like sequences, proteins, enzymes, bases, 3-D structures, etc.

- Development of a portal or new application is difficult because each internal data silo or application uses a different naming scheme to identify biological data items.

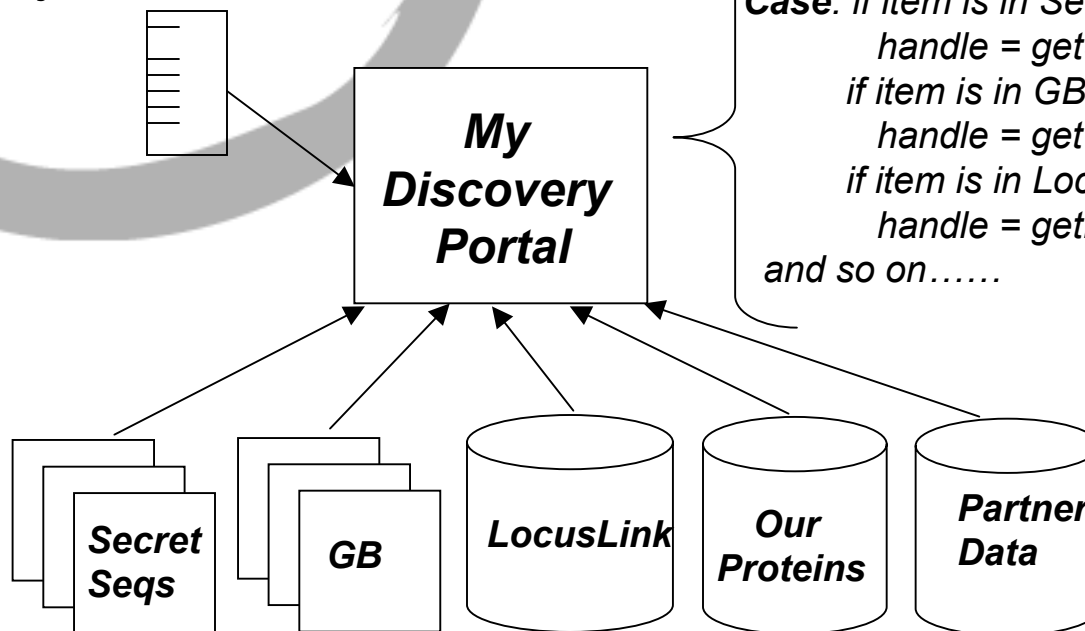
- Wide Area Data Distribution & Data Currency

Problem: A company obtains data from hundreds of public and private sources and wants to make it available to remote satellite offices so that all researchers always have the most up-to-date data. Today, all external data feeds or deliveries are received at their central office. Remote administrators periodically check a central FTP site to see if new data updates are available, but frequently fall behind.

- This causes unexplainable variances in the results generated by different research groups and rework.
- Research results based on out-of-date data are not as high quality as they can be, potentially lengthening the discovery process.

Intranet Portals and the need for Uniform Naming

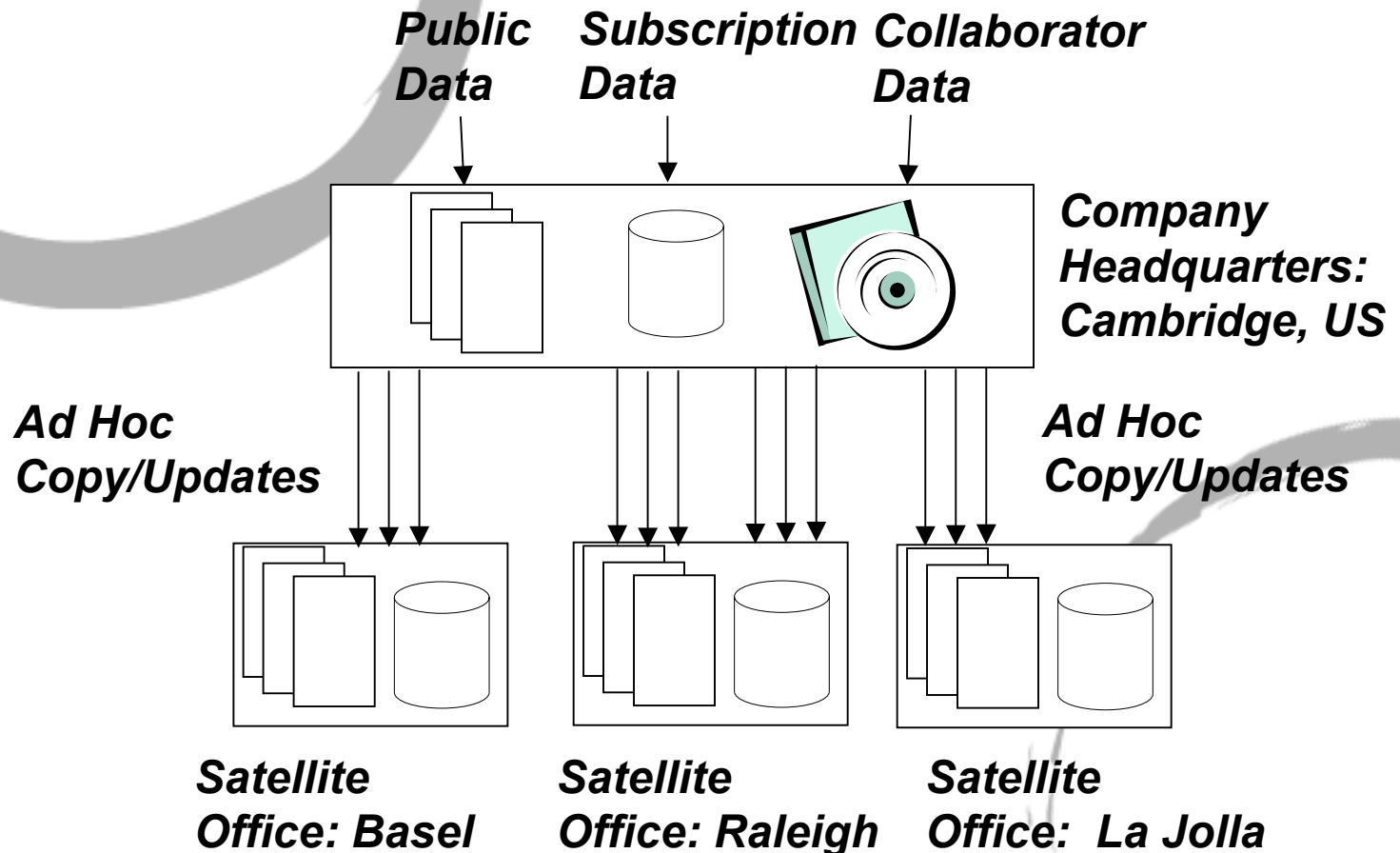
My bookmarks



Case: if item is in SecretSeqs:
 handle = getSecretSeqs(item);
if item is in GB:
 handle = getGB(item);
if item is in LocusLink:
 handle = getLocusLink(item);
and so on.....

Multi-site, Independent Files Systems, Relational Database, and N-Tier applications.

Wide Area Data Distribution and Data Currency



LSID Requirements

1. Scalability: LSID must scale to a very large number of items distributed across public or private wide area networks.
2. Multiple organizations: Biologically significant data objects named with LSID may be managed, owned and/or curated by multiple organizations. These organizations may be geographically separated.
3. Location Independence: Over time, a data item named with LSID may be migrated to a different network address, a different computer or a different administrative domain. This should not impact applications that refer to this data item.
4. Persistent names: The LSID name for a data item must be immutable and not change for the lifetime of that data item. The lifetime may include events such as data item migration, data store partitioning or network restructuring.
5. Replication and caching: For performance reasons, it must be possible to replicate or cache data items named with LSID.
6. Transparency: LSID names must be robust and transparent with respect to data item replication, migration and failure.
7. Rebind: LSID bindings can become stale, if the physical endpoint of an LSID changes. It must be possible to rebind an LSID.
8. Extensibility: LSID names must be extensible to support future revisions of the standard. Newer implementations of LSID must be able to resolve legacy name.

LSID Identifier Format

- **An LSID is represented as a Uniform Resource Name (URN) with the following format.**
 - URN:LSID:<Authority>:<Namespace>:<ObjectID>:<Version>
- **Examples:**
 - URN:LSID:ebi.ac.uk:SWISS-PROT/accession:P34355:3
 - URN:LSID:rcsb.org:PDB:1D4X:22
 - URN:LSID:ncbi.nlm.nih.gov:GenBank/accession:NT_001063:2

LSID Field Definitions

- <Authority>
 - The name of the organization that has defined the entity.
- <Namespace>
 - One or more statements that constrain the scope in which this identifier is evaluated.
- <ObjectID>
 - An alphanumeric sequence that uniquely defines this object within this namespace, as defined by the authority.
- <Version>
 - An optional field containing a unique integer that represents the version of the ObjectID. By convention, higher version numbers are more recent than lower version numbers.

URNs Provide Location Independence and Persistence

- LSIDs are encoded as a Universal Resource Names (URNs).
- A URN is an Internet resource with a name that has persistent significance and location independence.
 - The user of the URN can expect that someone else (or a program) will be able to find the resource.
 - The exact location of the Internet resource may change from time to time.
- In contrast, Universal Resource Identifiers (URIs) are used to name specific endpoints.
 - URIs are not persistent or location independent.

Identifier Semantics

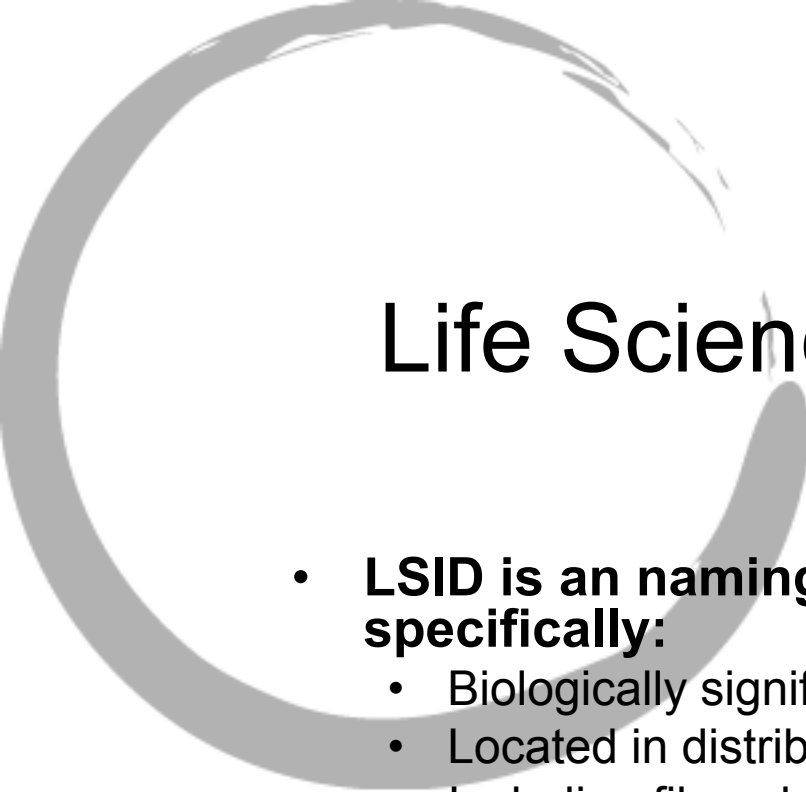
- LSIDs are opaque identifiers that are designed to be used by programs not people.
 - “The human mind seems to rebel at opaque identifiers, but they are necessary for creating reliable systems.”
- *Brian King, Sun Microsystems*
- User code should not infer semantic meaning about the objects that an LSID identifies.
- There is no way to add fields to the identifier.

LSID Use Cases Revisited

- Intranet Portal and Uniform Naming
 - LSID eliminates the need for 'case statement' conversion of multiple naming formats.
 - Simplifies portal integration, and more broadly point to point data silo integration in discovery pipelines.
 - Uniform naming format enables more complex models and ontologies, where data components and relationship definitions are physically remote or distributed.
- Wide Area, Data Distribution and Data Currency
 - Ensures that current data is available at all remote locations, on-demand.
 - Researchers always have the latest, best quality data sets.
 - Remote caches or replicas can determine which data items are out of date and update them automatically, in advance or on demand.
 - Version numbering facilitates reproducibility of experimental results.
 - Simplifies multi-organizational data sharing.

How LSID Relates to other Identity Standards

- Two types of identity
 - User identity, authentication, and privacy
 - Liberty Alliance
 - SAML
 - Microsoft Passport
 - GSS-API
 - Data identity
 - LSID (derived from SGNP)
 - SGNP (Secure Grid Name Protocol) submitted by Avaki to the OGSA (Open Grid Services Architecture) sponsored by IBM and Globus, within the GGF (Global Grid Foundation).
 - OIDs used in LDAP/ASN.1 based systems (compact & hierarchical)



Conclusion: Life Sciences Identifier – The Big Ideas

- **LSID is an naming standard for distributed data, specifically:**
 - Biologically significant data items,
 - Located in distributed data stores,
 - Including files, database records, and data objects managed by N-tier applications,
 - That are accessible over public and/or private networks,
 - And owned, managed, and/or curated by different academic, research, government or commercial organizations.
- **LSID names are semantically void/opaque with respect to the objects they identify.**
- **LSID replaces physical addresses with opaque, location independent identifiers expressed as URNs.**
- **LSID complements Web services standards such as SOAP/XML, WSDL, UDDI, SAML, WS-Security, MS Passport, Liberty Alliance, GSS-API and OGSA.**

LSID Standard Submission Status

- September, 2002: Submission of LSID specification to I3C Board of Directors for standardization vote.
- October 2004: Submitted to OMG through their standardization pipeline
- Today: Submitted and approved by OMG with implementations in Java, Perl, and C++